ПЕРСПЕКТИВЫ РАЗРАБОТОК ИСКУССТВЕННОГО МОРАЛЬНОГО АГЕНТА В КАЧЕСТВЕ ЭТИЧЕСКОГО СОВЕТНИКА

В. Ю. Перов
Санкт-Петербургский государственный университет
Санкт-Петербург

В исследовании рассматривается возможность создания Искусственных Моральных Агентов (ИМА) на основе алгоритмов Искусственного Интеллекта (ИИ), которые способны работать на основании максимально широкого спектра этических знаний и могут быть использованы в качестве Этических Советников (ЭС). Выделены несколько потенциальных направлений исследований и разработок. (1) Создание ИМА ЭС на основе моральных ценностей российского общества. (2) Разработка «персонифицированного» ИМА ЭС. (3) Создание ИМА ЭС профессиональной направленности. (4) Создание ИМА ЭС в виде программ «этического надзора» для технологий на основе алгоритмов ИИ. В ходе исследования были сформулированы основные преимущества и этические риски для этих видов ИМА ЭС.

Ключевые слова: искусственный интеллект, искусственный моральный агент, этический советник

PROSPECTS FOR THE DEVELOPMENT OF ARTIFICIAL MORAL AGENT AS ETHICAL ADVISOR

V. Y. Perov
Saint-Petersburg State University
Saint-Petersburg

The study considers the possibility of creating Artificial Moral Agents (AMA) based on Artificial Intelligence (AI) algorithms that are able to work on the basis of the widest possible range of ethical knowledge and can be used as Ethical Advisors (EA). Several potential areas of research and development have been identified. (1) Creation of the AMA EA based on the moral values of the Russian society. (2) Development of a "personalized" AMA EA. (3) Creation of AMA EA of professional orientation. (4) Creation of AMA EA in the form of "ethical supervision" programs for technologies based on AI algorithms. The study formulated the main benefits and ethical risks for these types of AMA EA.

Keywords: artificial intelligence, artificial moral agent, ethical advisor

На протяжении большей части истории развития человечества одним из наиболее важных морально-практических вопросов была проблема формирования нравственно правильного морального агента, то есть людей способных к добродетельным поступкам, которые могли бы быть достойными членами хорошего общества. Это обстоятельство обусловило то, что начиная с античности проблема места и роли знания в моральном воспитании и поведении стала одной из центральных для теоретической этики. Конечно, оценка значения этического знания в обеспечении добродетельного поведения разнилась у отдельных мыслителей. Так, традиция «этического гносеологизма», уходящая корнями в учения Сократа и Платона, зачастую отождествляет этичность (добродетельность) со знанием моральных явлений, полагая, что последнее является не только необходимым, но и достаточным условием для нравственно правильного поведения. Такая крайняя точка зрения была подвергнута сомнению и критике уже в трудах Аристотеля, который различал мыслительные (дианоэтические) и нравственные (этические) добродетели. При этом он не только не отрицал возможность и важность этического знания, но и отстаивал его необходимость для сознательного выбора. В концепции Аристотеля для фронезиса как обязательного условия для добродетельной и хорошей жизни необходимо знание о том, что такое добродетели, каковы особенности добродетельного характера действующего лица и в каких условиях совершается добродетельный поступок. Это знание рассматривалось им не как теоретическое (созерцательное), а как практическое знание, которое должно включать жизненный опыт самого человека и знание о нравственном опыте других людей. При всей привлекательности идей Аристотеля возникает существенный вопрос/ проблема: насколько способны люди овладеть столь большим объемом знаний? А за прошедшее после его жизни время накопились огромные знания как в области этики, так и в отношении нравственного (позитивного и негативного) опыта человечества. Кроме того, остается вопрос о том, как можно не только получать, но и учитывать в принятии

моральных решений знания об особенностях личностных характеристик действующих моральных агентов и наличных обстоятельствах. Создается устойчивое впечатление, что такое знание является необъятным и находится за разумными пределами человеческих способностей.

Ситуация, возможно, изменяется в связи со становлением информационного (цифрового) общества. Широкое публичное распространение технологий, особенно основанных на алгоритмах искусственного интеллекта (ИИ), актуализируют многочисленные этические проблемы, формируя такое особое направление как киберэтика, включающую «машинную мораль» или «мораль машин», которая может рассматриваться в различных смыслах. С одной стороны, под «моралью машин» подразумевается наделение алгоритмов ИИ какими-то определёнными нравственными правилами, предполагающими возможности принятия ими этически значимых решений. Наиболее наглядными примерами в этом контексте являются исследования в отношении беспилотных транспортных средств. С другой стороны, появление голосовых помощников, чат-ботов и т.д. побуждает людей использовать их для поиска ответов на вопросы во многих областях, в том числе и в сфере нравственности. IT-отрасль не может игнорировать запросы подобного рода, в связи с чем, вполне оправданной выглядит идея о необходимости создания «этических помощников» или «этических консультантов» на основе алгоритмов ИИ. Краткая суть обсуждаемых идей состоит в следующем: созданный на основе алгоритмов ИИ Искусственный Моральный Агент (ИМА) будет своеобразным воплощением этической мудрости в области моральной философии за всю ее историю, обладателем знаний о моральном опыте человечества, источником информации о существующих в современных обществах моральных предпочтениях людей различных стран и культур и т.д. Трудно представить, что подобный объем знаний может быть доступен какому-то отдельному человеку. Кроме того, к прогнозируемым преимуществам следует отнести потенциальную беспристрастность выносимых решений на основе обработки упомянутой информации. Но самым главным преимуществом предполагается то обстоятельство, что создаваемый ИМА будет выступать не в роли «вершителя судеб» и «глашатая моральных истин в последней инстанции», а только в роли Этического Советника (ЭС). Иными словами, ИМА ЭС не подменяет самостоятельности морального выбора людьми, а выступает в качестве «этического помощника».

Примером практической попытки создать такого ИМА ЭС является проект Ask Delphi (Allen Institute for AI). На сайте представлено следующее его описание: «Delphi — исследовательский прототип, предназначенный для моделирования моральных суждений людей в различных повседневных ситуациях. Эта демонстрационная версия показывает возможности и ограничения современных моделей» [1]. Среди наиболее существенных ограничений данного проекта можно упомянуть упрощенность возможных вопросов и ещё большую простоту ответов, а также их языковую и национально- культурную предвзятость, что признается самими разработчиками. В частности, на сайте в разделе FAQ сформулирован утвердительный ответ на вопрос: «Отражает ли Delphi в основном ориентированную на США культуру и моральные ценности?»: «Короткий ответ: Да. Delphi обучается на базе Банка Норм Здравого Смысла (Commonsense Norm Bank), который содержит суждения американских краудсорсеров, основанные на ситуациях, описанных на английском языке. Вероятно, это отражает то, что вы бы назвали группами “большинства” в США, то есть белыми, гетеросексуальными, трудоспособными, имеющими жилье и т.д. Поэтому не ожидается, что он будет отражать какой-либо другой набор социальных норм. Тем не менее, как ни странно, он все же может уловить некоторые культурные различия (примеры см. в статье). Но необходимо проделать гораздо больше работы, чтобы рассказать Delphi о разных культурах, разных странах и разных подгруппах в США» [1]. Несмотря на существенные недостатки, опубликованные результаты и анализ работы данного ресурса позволяют по-новому взглянуть на плодотворные перспективы реальных и потенциальных ИМА ЭС, основные направления разработки которых можно сформулировать следующим образом.

(1) Создание ИМА ЭС, учитывающих особенности нравственных ценностей не только США, но и других стран и культур, например, российского общества. Следует отметить, что на этом пути велики шансы повторения опыта разработчиков Ask Delphi в плане получения предвзятого ИМА ЭС, только пока не очень понятно, этические предпочтения какого именно морального агента окажутся в приоритете, учитывая многонациональность и поликультурность российских реалий. Учесть же алгоритмически все многообразие существующих и формально неопределенных моральных норм и ценностей, включая особенности их языкового выражения, пока не представляется возможным.

(2) Разработка «персонифицированного» ИМА ЭС. Суть этой идеи заключается в том, чтобы создаваемый ресурс при формулировании предлагаемых этических решений использовал не только общеэтические знания и информацию об общепринятых нормах нравственности, но и моральные убеждения пользователя. В обобщенном виде это могло бы быть обеспечено (а) пользовательскими настройками и (б) на основе анализа при помощи алгоритмов ИИ прошлых моральных предпочтений (по аналогии с системами интеллектуального поиска и таргетированной рекламы). Но и потенциально возникающие в этом случае проблемы и этические риски будут аналогичными. Например, велика вероятность появления чего-то вроде «пузыря моральных фильтров».

(3) Создание ИМА ЭС профессиональной направленности на основе существующих этических кодексов и практик. Преимущество данного направления состоит в том, что предполагаемые ИМА ЭС будут изначально основаны на конечном перечне моральных норм и ценностей (сформулированных в этических кодексах профессий или организаций), определенности базовых этических практик и адресной моральной агентности. Подобные ресурсы в виде «виртуальных этических помощников» могут помочь принятию морально значимых решений в сложных ситуациях профессиональной деятельности и использоваться для работы этических комитетов и комиссий.

(4) Возможно, что наиболее перспективным направлением на основе современного состояния цифровых технологий – это создание ИМА ЭС для самих алгоритмов ИИ. Речь идет о программах своеобразного «этического надзора» или этического аудита алгоритмов ИИ. Для этого могут быть использованы рекомендации, документы и «этические стандарты» серии P7000 Института Инженеров электротехники и Электроники (IEEE). Кроме того, в качестве источника моральной нормативности можно опираться на положения «Кодекса этики в сфере ИИ» (2021), «Кодекса этики использования данных» (2018), Хартии «Цифровая этика детства» (2022), а так же, целого ряда имеющихся в сфере этики цифровых технологий отечественных, зарубежных и международных разработок. Использование ИМА ЭС «этического надзора» поможет обеспечить прозрачность разрабатываемых алгоритмов ИИ и будет способствует повышению доверия к ним.

Выявленные возможные направления создания Искусственного Морального Агента (ИМА) в качестве Этического Советника (ЭС) могут выглядеть довольно фантастическими в настоящее время. Более того, есть большая степень вероятности, что не найдут своего воплощения в конкретных технологиях в полной мере. Но можно с уверенностью утверждать, что такие исследования и разработки существенно обогатят как опыт работы с алгоритмами ИИ, так и внесут существенный вклад в развитие этической теории, прикладных и профессиональных этик.

Исследование выполнено за счет гранта Российского научного фонда № 22-28-00379 «Трансформации морального агентства: этико-философский анализ» (https://rscf.ru/project/22-28-00379/).

ЛИТЕРАТУРА

1. Ask Delphi. Allen Institute for AI. URL: https://delphi.allenai.org/ (дата обращения: 26.03.2023).